Published in: Draxler, Christoph/Kleber, Felicitas (eds.): 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum. 12. - 14. Oktober 2016, München, Deutschland. München: Institut für Phonetik und Sprachverarbeitung, Universität München, 2016. pp. 9-11.

The Karl Eberhards Corpus of spontaneously spoken southern German in dialogues – audio and articulatory recordings

Denis Arnold, Fabian Tomaschek

University of Tübingen, Germany

denis.arnold@uni-tuebingen.de, fabian.tomaschek@uni-tuebingen.de

Abstract

The current paper presents a corpus containing 35 dialogues of spontaneously spoken southern German, including half an hour of articulography for 13 of the speakers. Speakers were seated in separate recording chambers, mimicking a telephone call, and recorded on individual audio channels. The corpus provides manually corrected word boundaries and automatically aligned segment boundaries. Annotations are provided in the Praat format. In addition to audio recordings, speakers filled out a detailed questionnaire, assessing among others their audio-visual consumption habits.

Index Terms: corpus, spontaneous speech, conversation, articulography, German.

The authors contributed equally to the paper and its content.

1. Introduction

Recently, Wagner, Trouvain and Zimmerer [11] have shown that phonetic studies mainly rely on 'scripted' speaking styles, i.e. speech which is recorded in a highly controlled environment in the phonetic lab ($\approx$ 70%). By contrast, 'unscripted' speaking styles i.e. styles without any a priori control, are in the minority ($\approx$ 15%). This is not only the case in acoustic analyses but also in articulatory analyses [10, 9, 12, 13]. This imbalance in speaking styles affects our models about speech production, as large amounts of variations in casual speaking situations are neglected. One possibility to overcome this shortcoming is to record corpora of spontaneously spoken language.

There are plenty of corpora which contain spontaneously elicited speech (see for a extensive list [2]). However, to our knowledge, none of them contain spontaneously elicited speech in dialogues, as is the case in the Kiel Corpus [4, 5] and the GECO corpus [7]. In the Kiel Corpus, spontaneous speech was elicited by mimicking a telephone call as well as discussing non-matching videos with a dialogue partner; in the GECO corpus, speakers talked spontaneously about different topics. However, they did not know each other before the conversation.

The current paper presents the Karl Eberhards Corpus (KEC) of spontaneously spoken southern German elicited in dialogues. Dialogue partners were not instructed on the topic of their conversation, nor did the experimenters interfere with the speakers during recording. In contrast to the GECO corpus, dialogue partners were well acquainted friends. In addition, participants had a conversation of one hour.

At the time of publication, we have recorded 35 one hour long dialogues between two speakers. In addition to pure audio recordings, the corpus contains 13 speakers for which half hour long recordings of articulography were recorded, amounting to roughly 2 hours of speech without pauses. Finally, all speakers provided detailed personal information in the form of a questionnaire.

2. Recordings

We targeted speakers in their mid twenties and early thirties. Most of the speakers were students of the University of Tübingen. We recorded 11 male and 28 female speakers. Their median age was 25 years, with a range from 19 to 33 years. Speakers were compensated for participation either by course credits or €10/hour. We insisted that speakers could only take part in the recording if they were well acquainted with their partners and frequently spoke with him. In this way we ensured that 1. speakers were capable to chat for at least an hour and 2. that they found a common topic to discuss.

2.1. Audio

Recordings were performed in two separated sound-treated recording booths at the *Seminar für Sprachwissenschaften in Tübingen* from 2014 till 2016 and are still going on. During the one hour of recording, speakers were not interrupted. Every speaker was recorded onto an individual audio channel, allowing the investigation of e.g. interruptions and turn taking. The format of the recording for every speaker is:

- Six ten minutes long wave files in the *.wav format (Sampling rate: 44100 Hz, 32-bit float; in case of articulography: 22050 Hz, 16-bit float).
- Six Praat TextGrids for the respective wave files (UTF-8 Encoding, [1]).
- One questionnaire in *.txt format (UTF-8 Encoding)

2.2. Articulography

In a portion of the dialogues we recorded electromagnetic articulography of one of the speakers for one half of an hour. Articulatory recordings were performed by means of the NDI wave articulograph at a sample rate of 400 Hz. The audio sample rate was 22050 Hz, 32-bit float. Figure 1 illustrates the recorded sensor locations on tongue back (TB), tongue mid (TM), tongue tip (TT), upper teeth (UT), lower teeth (LT), upper lip (UL), lower lip (LoL), left lip edge (LL), jaw (J). Apart from the jaw and LL sensor, all sensors were attached along the midsagittal plane. We used three head positions (nasion (N), left/right mastoid (LM/RM)) as reference sensors for correction of head movements. We also recorded a bite plate recording in order to centralize sensor positions.

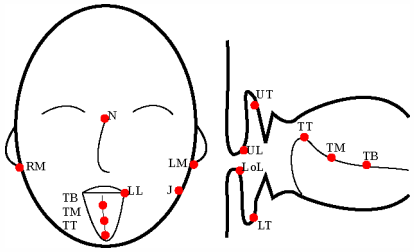

Figure 1: *Illustration of sensor positions. Left: frontal illustration. Right: midsagittal cut through the mouth. See section 2.2 for details on sensors.*

Table 1: *20 most common words and their absolute and relative frequencies.*

Word	Raw Freq.	Rel. Freq.	Word	Raw Freq.	Rel. Freq.
ja	10885	0.043	also	2859	0.011
ich	9650	0.038	der	2805	0.011
und	6507	0.026	halt	2700	0.011
so	6109	0.024	ist	2655	0.011
das	5586	0.022	nicht	2431	0.010
die	4968	0.020	du	2287	0.009
dann	4131	0.016	war	2034	0.008
auch	3874	0.015	was	1887	0.008
da	3155	0.013	hat	1845	0.007
aber	3123	0.012	'ne	1749	0.007

2.3. Annotation

Our focus was to provide precise annotations at the word level. For this, Praat TextGrids were corrected manually. Annotations at segment level were performed by means of a forced aligner [6] within the corrected word boundaries.

3. Questionnaire

After the recording, participants filled out a questionnaire. The questionnaire was in German. An English translation of the questionnaire is provided. Participants were allowed to skip questions, marking the skipped questions by `keine Angabe` (*not specified*). We assessed answers to the following multiple types of questions, among others:

- **General questions:** Gender, year of birth; educational level; occupation; life situation;
- **Linguistic development:** Native language, proportion of language use.
- **Reading habits**.
- **Consumption** of audio-visual media.

4. Statistics

Table 3 compares the KEC in its current, preprocessed and annotated form with two existing corpora. The statistics for the Kiel Corpus cover the recordings of spontaneously spoken speech. KEC and GECO have roughly the same total number of words. But all corpora differ with respect to their number of unique and consequently in their total/unique word ratio, indicating how "often" a single token was used. In both, KEC and Kiel every token was used roughly 15 times, in the GECO it was used roughly 20 times. The corpora differ with respect of the number of words per minute, which can be regarded to be representative of average speaking rate. The duration of the

Table 2: *20 most common words in the SDEWAC corpus.*

Word	Rel. Freq.	Word	Rel. Freq.
die	0.037	des	0.008
der	0.035	nicht	0.008
und	0.029	für	0.008
in	0.018	auf	0.008
den	0.012	im	0.008
zu	0.011	sich	0.008
das	0.011	ein	0.007
von	0.010	eine	0.007
ist	0.009	es	0.007
mit	0.009	sie	0.007

Table 3: *Corpus statistics. See section 4 for details.*

	KEC	Kiel	GECO
Total words	240299	37257	246621
Unique tokens	15783	2241	12349
Ratio total/unique words	15.2	16.6	19.9
Duration in min.	996.5	214.4	1163.1
Words/min	241	174	212
% rare words	4.8	5.7	4.4

corpora was calculated by excluding the pauses. The Kiel Corpus has the lowest and the KEC has the highest. Furthermore, we calculated the percentage of rare words (frequency of occurrence in corpus < 10). In all the three corpora they are $\sim 5\%$ of the total number of words.

Table 1 shows the 20 most frequent words in the corpus. Note that *ja* is the most common word in the KEC, which is also the case in the GECO and the Kiel Corpus. This is especially striking when comparing these frequencies to frequencies in written corpora, e.g. like SDEWAC [3, 8] (cf. Table 2). Moreover, *'ne* is more frequent than its canonical form *eine* (ind. article, fem.), which is not present at all among the most frequent 20 words. Furthermore, the corpus contains ~6050 hesitations and ~3655 laughs, which is interesting for researchers of hesitations and interruptions [14].

5. Distribution

The KEC is planned to be submitted to Clarin-D [2]. In addition to wave files, articulography recordings and Praat TextGrids, the current distribution of the KEC contains R scripts to process the corpus, such as reading in Praat TextGrids, reading in articulography, tagging articulography. Furthermore, a lexicon of frequently reduced forms and their canonical equivalents is provided.

6. Acknowledgments

The authors thank Deniz Cevher, Jan Hoffmann, Ronaldo Rodrigues, Gina Hermann, Mareike Vermehren, Rachel Dockweiler, Verena Heusser and Jessica Viertel for their help creating the corpus. The paper and the corpus were funded by the Alexander von Humboldt Chair awarded to R. H. Baayen.

References

[1] Paul Boersma and David Weenink. *Praat: doing phonetics by computer [Computer program], Version 5.3.41, retrieved from http://www.praat.org/.*

[2] *Clarin-D*. URL: http://www.clarin-d.de/.

[3] Gertrud Faaß and Kerstin Eckart. "SdeWaC - A Corpus of Parsable Sentences from the Web". In: *Language Processing and Knowledge in the Web*. Ed. by Iryna Gurevych, Chriss Biemann, and Torsten Zesch. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 61–68.

[4] Klaus J. Kohler. *Labelled data bank of spoken standard German – The Kiel Corpus of read/spontaneous speech.*

[5] Benno Peters. *Die Datenbasis The Kiel Corpus.*

[6] S. Rapp. "Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German". In: *Proceedings of ELSNET goes east and IMACS Workshop*. Moscow, 1995.

[7] A. Schweitzer and N. Lewandowski. "Convergence of Articulation Rate in Spontaneous Speech". In: *Proceedings of Interspeech 2013*. Lyon, 2013.

[8] Cyrus Shaoul and Fabian Tomaschek. *A phonological database based on CELEX and N-gram frequencies from the SDEWAC corpus*. 2013. URL: https://fabiantomaschek.files.wordpress.com/2016/07/tomaschek%5C_corpus%5C_readme.pdf.

[9] Fabian Tomaschek et al. "Vowel articulation affected by word frequency". In: *Proceedings of the 10th ISSP*. Cologne, 2014.

[10] Fabian Tomaschek et al. "Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience". In: *Proceedings of the Interspeech*. Lyon, 2013.

[11] Petra Wagner, Jürgen Trouvain, and Frank Zimmerer. "In defense of stylistic diversity in speech research". In: *Journal of Phonetics* 48 (2015), pp. 1–12.

[12] Martijn Wieling et al. "Investigating dialectal differences using articulography". In: *Proceedings of the 18th ICPHS*. Glasgow, 2015.

[13] Martijn Wieling et al. "Investigating dialectal differences using articulography". In: *Journal of Phonetics* (submitted).

[14] Martijn Wieling et al. "Variation and change in the use of hesitation markers in Germanic languages". In: *Language Dynamics and Change* (2016).